Kaggle Competition Overview

Ruslan Khalitov

ruslan.khalitov.edu@gmail.com

January 2020

Overview

1. Kaggle Platform Overview
2. My Kaggle account
3. Google Landmark Recognition Challenge

Kaggle Platform Overview

Kaggle is an online community of data scientists and machine learning practitioners with over than 130 000 participants worldwide.

Competition Formats:

1. Simple Competitions
2. Two-stage Competitions
3. Code Competitions

My Kaggle account

Ruslan Khalitov

Data Scientist at DataDuck

Limassol, Cyprus

Joined 3 years ago · last seen in the past day

Followers 2

Following 2

Competitions Expert

Home | Competitions (20) | Datasets | Kernels | Discussion (1) | •••

Edit Profile

Competitions Summary

Competitions Expert

Current Rank	Highest Rank
788 of 130,319	711

0	1	3

Competitions: 20
Solo: 19 (95%)
Team: 1 (5%)

My Kaggle account

Santander Customer Transaction Prediction
Can you identify who will make a transaction?
Featured · 10 months ago · banking, tabular data, binary classification

221/8802
Top 3%

Google Landmark Recognition 2019
Label famous (and not-so-famous) landmarks in images
Research · 8 months ago

61/281
Top 22%

Generative Dog Images
Experiment with creating puppy pics
Research · Code Competition · 5 months ago

73/927
Top 8%

IEEE-CIS Fraud Detection
Can you detect fraud from customer transactions?
Research · 4 months ago · tabular data, binary classification

612/6381
Top 10%

Google Landmark Recognition Challenge overview

The source code repository:
https://github.com/RuslanKhalitov/google_landmark_challenge_2019

Parameter	**Specification**
Competitition type	Code Competition
Objective	Landmark Recognition
Problem type	Multiclass Classification
Number of images	$Train \sim 4M$, $Test \sim 400K$
Images size	$224x224$
Total size	$\sim 500GB$
Number of classes	200 000
Timeline	3 months
Hardware	Personal instance with 1 x Nvidia Tesla V100

Solution Pipeline

- Metric analysis
- Dataset downloading
- Data cleaning
- Image resizing
- Data augmentation
- Learning a pre-trained model
- Inference
- Stacking
- Submission

Evaluation Metric

Submissions were evaluated using Global Average Precision (GAP):

$$GAP = \frac{1}{M} \sum_{i=1}^{N} P(i) \cdot rel(i)$$

where:

N — total number of predictions returned by the system, across all queries
M — total number of queries with at least one landmark from the training set visible in it
$P(i)$ — precision at rank i
$rel(i)$ — relevance of prediction i: it's 1 if the i-th prediction is correct, and 0 otherwise

Data Cleaning

Data Cleaning

Three stages of the dataset cleaning:

1. Remove any places that are not in target set, but recognized by pre-trained VGG16 Places365
2. Remove portraits and selfies using pre-trained Faster R-CNN from the TorchVision library
3. Remove random images (cats, helicopters, plants, etc.) using the pre-trained ImageNet classifier with ResNet-50

Data Augmentation

Augmentations applied to images was the most important step before the learning process.

Image transformation set:

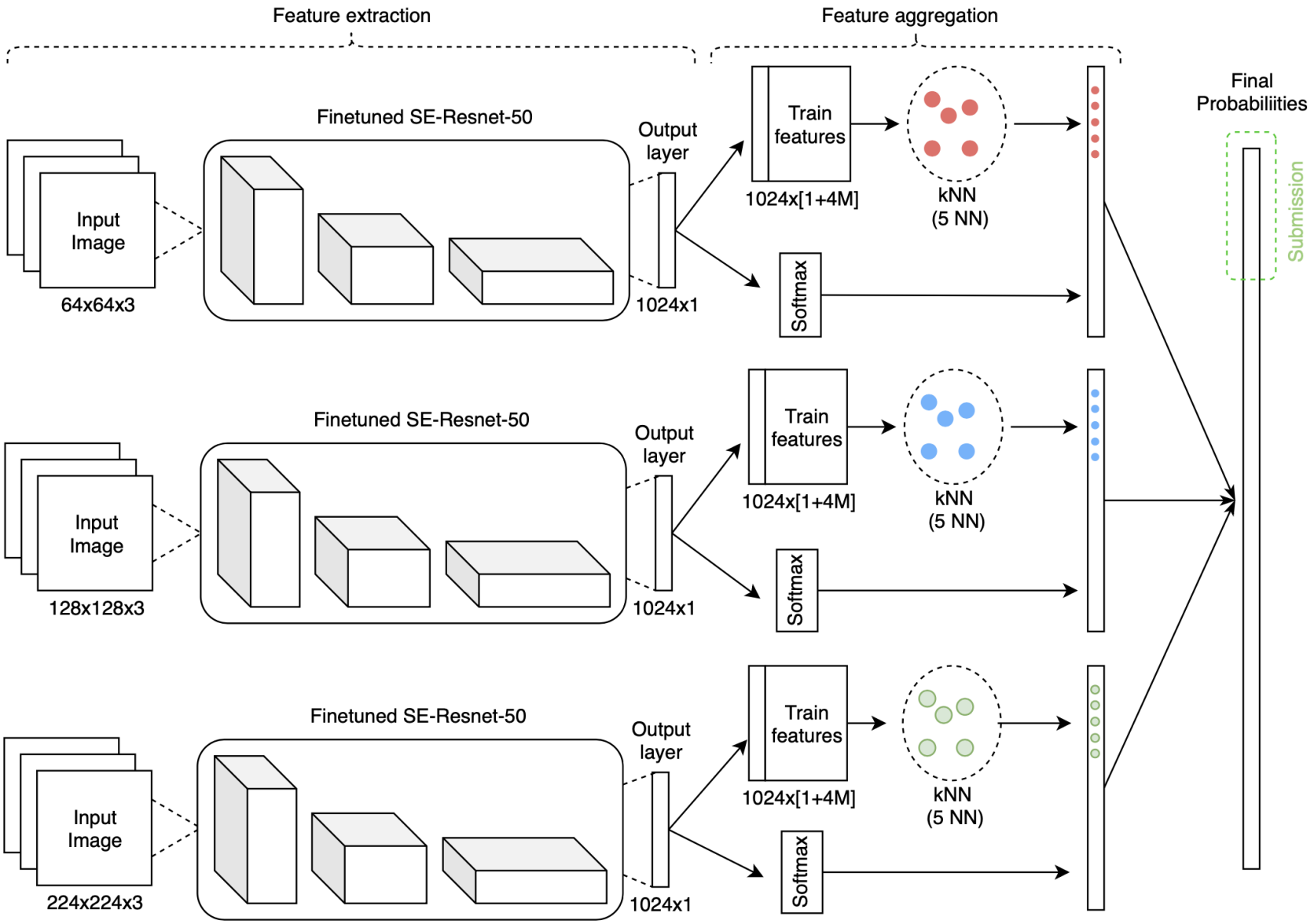
1. Color Normalization
2. Random Resized Crop
3. Random Affine Transformation
4. Horizontal Flip
5. slight Grid Distortion

Training process description

Parameter	Specification
Main Model	SE-ResNet50
Images Size	[64x64, 128x128, 224x224]
Batch Size	[512, 128, 64]
Optimizer	Adam
Additional LR Scheduler	Implemented SGDR from the original paper[1]
Activation Function	SoftMax
Loss	Cross entropy loss
Average experiment length	5 days
Postprocessing model	kNN
Meta-algorithm	Weighted Average

[1] I.Loshchilov and F.Hutter. "SGDR: Stochastic Gradient Descent With Warm Restarts". ICLR 2017. arXiv preprint arXiv:1608.03983.

Inference process description



Thank you!